Social Network Analysis and Mining (2018) 8:18
https://doi.org/10.1007/s13278-018-0496-z

ORIGINAL ARTICLE

BullyBlocker: toward an interdisciplinary approach to identify cyberbullying

Yasin N. Silva[1] · Deborah L. Hall[1] · Christopher Rich[1]

Received: 13 December 2016 / Revised: 20 February 2018 / Accepted: 1 March 2018

Abstract
Cyberbullying is the deliberate use of online digital media to communicate false, embarrassing, or hostile information about another person. It is the most common online risk for adolescents, yet well over half of young people do not tell their parents when it occurs. While there have been many studies about the nature and prevalence of cyberbullying, there have been relatively few in the area of automated identification of cyberbullying that integrate findings from computer science and psychology. The goal of our work is thus to adopt an interdisciplinary approach to develop an automated model for identifying and measuring the degree of cyberbullying in social networking sites, and a Facebook app, built on this model, that notifies parents about the likelihood that their adolescent is a cyberbullying victim. This paper describes the challenges associated with building a computer model for cyberbullying identification, presents key results from psychology research that can be used to inform such a model, introduces a holistic model and mobile app design for cyberbullying identification, presents a novel evaluation framework for assessing the effectiveness of the identification model, and highlights crucial areas of future work. Importantly, the proposed model—which can be applied to other social networking sites—is the first that we know of to bridge computer science and psychology to address this timely problem.

Keywords Cyberbullying · Automated identification · Social networks · Facebook · Psychology · Cyberbullying factors · Vulnerability factors

1 Introduction

Over half of adolescents have been bullied online, about the same number have engaged in cyberbullying, and more than one in three young people have experienced cyber-threats online (http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html). Cyberbullying can take multiple forms such as posting hurtful or threatening messages online, spreading rumors on social networking sites, taking and posting unflattering pictures of a person, or circulating sexually suggestive pictures or messages about a person. The consequences of cyberbullying—which can include anxiety, depression, and even suicide (Van Geel et al. 2014)—are detrimental on both an individual and a societal level. Despite the growing prevalence of cyberbullying, well over half of young people do not tell their parents when cyberbullying occurs (http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html). Moreover, while there have been many studies about the nature and prevalence of cyberbullying and even a few on cyberbullying measures for mobile and chat-based venues, there has been little work on the design and implementation of automated models and tools to identify cyberbullying that bridges findings from computer science and psychology.

The closest relevant work, which uses machine learning-based models to identify cyberbullying in social networking sites, e.g., (Dinakar et al. 2011; Rafiq et al. 2015; Homsein-mardi et al. 2016; Reynolds et al. 2011; Huang et al. 2014; Squicciarini et al. 2015), has several crucial limitations. For instance, these previously proposed models focus on

This work was supported by National Science Foundation Award # 1719722, the Dion Initiative for Child Well-Being and Bullying Prevention and ASU NCUIRE awards.

✉ Yasin N. Silva
ysilva@asu.edu

Deborah L. Hall
d.hall@asu.edu

Christopher Rich
cdrich2@asu.edu

[1] Arizona State University, Glendale, AZ, USA

Published online: 14 March 2018

the identification of cyberbullying in a single message or picture. While insults may appear in single messages, the models do not consider that in many cases cyberbullying occurs as a sequence of insulting or harassing events. The accuracy and reliability of a model for identifying cyberbullying will likely increase to the extent that the number and frequency of insulting messages are taken into account. Moreover, for the most part, previous work has not integrated critical findings from psychology research on the nature of cyberbullying, including risk factors and negative outcomes of cyberbullying and patterns of cyberbullying over time. Additionally, to our knowledge, previous work has yet to address key related problems like the automated generation of a list of anti-bullying resources for parents (e.g., websites, hotlines) based on specific characteristics of the cyberbullying that an adolescent is experiencing. Finally, few of the previous papers have sought to integrate a cyberbullying identification model into an actual app that can help parents and potential victims.

The goal of our work is thus to adopt an interdisciplinary framework to study, design, and implement a model to identify cyberbullying among adolescents in social networking sites. The initial model has been used to build BullyBlocker, a mobile app that identifies cyberbullying on Facebook and generates a customized list of anti-bullying resources for parents. BullyBlocker alerts parents of potential cyberbullying instances by providing them with a Bullying Rank (BR) that estimates the probability that their child is being bullied, allowing them to identify and address this form of online aggression. While Facebook is the most common social media platform for teens (http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/), the principles and design used in BullyBlocker can also be applied to other social networking platforms. Furthermore, similar models could also be used to identify a broad range of negative outcomes that may result from or be exacerbated by cyberbullying, such as depression, substance use, or self-destructive behavior. The primary contributions of this paper are:

- The design of a holistic model to identify cyberbullying that builds on previous research findings on cyberbullying in adolescents. The proposed model integrates crucial findings from psychological research on predictors of cyberbullying, as well as the relative strength of various predictors and temporal aspects of cyberbullying. Furthermore, rather than focusing on cyberbullying prediction or classification for a single message or picture, the proposed model considers streams or bursts of messages in conjunction with information from adolescents' social media profiles.
- The design and implementation of a mobile app (BullyBlocker) that uses the proposed model to identify cyberbullying in Facebook. We present the implemented app architecture and a discussion of the key software components.
- The introduction of a module that provides a customized list of parent/victim resources using information pertaining to the nature of specific cyberbullying instances.
- The development of an innovative framework for evaluating the accuracy of holistic cyberbullying identification models that involves a simulated social network with content from real-world cyberbullying interactions and a comparison of the results of automated identification models with human assessments of cyberbullying likelihood. We present the results of evaluating the BullyBlocker identification model using this framework.
- The identification of challenges and opportunities to integrate the latest results from psychology and social network data analysis to address a problem of great social impact. We also highlight areas that warrant further study within psychology.
- The public and no-cost availability of the BullyBlocker app (1.0) in the Apple App Store (https://itunes.apple.com/us/app/bullyblocker-app/id1236).
- The public availability of the source code of the evaluation framework as well as the real-world datasets that it uses to generate the social network interactions (https://bullyblocker.project.asu.edu/data).

This paper builds on and marks a significant extension of an abstract that appeared in (Silva et al. 2016). In particular, the current paper expands our previous work by integrating: (1) an extended identification model that includes (a) multiple new cyberbullying factors—including insulting video comments, race and ethnicity, frequency of internet use, past bullying experiences, sexual orientation, mental health history, disciplinary problems, and substance use, (b) the use of correlation coefficients (r) identified in meta-analytic reviews of cyberbullying research to increase the accuracy of the weights assigned to the different factors in the model, and (c) the use of psychology research to estimate the temporal parameters in our model; (2) a more detailed description of our cyberbullying identification model; (3) an innovative evaluation framework for holistic models that integrate profile features and message streams; (4) an extended architecture diagram, (5) an in-depth explanation of the motivation behind our project that highlights how the present research addresses several important limitations of previous work; (6) an expanded and more detailed review of previous machine learning-based cyberbullying identification models and relevant empirical findings in psychology; (7) the design guidelines for two additional identification models as future improvements; and (8) a detailed presentation of how additional results in psychology will be integrated into these identification models.

The remainder of this paper is organized as follows. Section 2 presents the background and related work, Sect. 3 describes the proposed model and app design guidelines, Sect. 4 discusses our novel evaluation framework for assessing the accuracy of holistic cyberbullying identification models and the results of our evaluation of the BullyBlocker identification model, Sect. 5 describes some key paths for future work, and Sect. 6 concludes the paper.

2 Background and related work

Prior contributions from computer science that propose models for identifying cyberbullying (e.g., Dinakar et al. 2011; Rafiq et al. 2015; Hosseinmardi et al. 2016; Reynolds et al. 2011; Huang et al. 2014; Squicciarini et al. 2015) rely primarily on machine learning classification or prediction models that analyze text features (e.g., comments and posts) (Dinakar et al. 2011; Reynolds et al. 2011), externally annotated images or videos (Rafiq et al. 2015), or both (Hosseinmardi et al. 2016). Yet, the accuracy of these methods, as highlighted in (Huang et al. 2014), remains limited. Moreover, a major drawback of cyberbullying research within computer science is that it has largely ignored relevant psychology research findings. That is, while the results of studies incorporating some social network features (Huang et al. 2014) and user demographics (Squicciarini et al. 2015) to improve accuracy are promising related efforts, a core open challenge is how to effectively integrate insights from psychology to improve automated identification models.

To develop an effective cyberbullying identification model, we are thus building on empirical work from within psychology. Although findings regarding the prevalence, determinants, and even the definition of cyberbullying vary somewhat within the psychology literature (Kowalski et al. 2014; Tokunaga 2010), examination of the cumulative findings, across multiple studies, reveals some important trends and emerging areas of agreement (Kowalski et al. 2014; Wolke et al. 2016; Patchin and Hinduja 2012; Guo 2016). For example, seeming inconsistencies across studies in prevalence rates of cyberbullying among different age groups have more recently shed light on a potential curvilinear relation between age and rates of cyberbullying in children and adolescents (Tokunaga 2010; Williams and Guerra 2007), with the prevalence of cyberbullying victimization at its highest during the later years of middle school (Williams and Guerra 2007). Findings with respect to gender and cyberbullying are similarly complex, with some studies indicating that adolescent girls experience higher rates of cyberbullying than adolescent boys (Ortega et al. 2009; Kowalski and Limber 2007; Ybarra and Mitchell 2008; Mishna et al. 2012) and others finding no systematic gender differences (Williams and Guerra 2007; Hinduja and Patchin 2008). There is greater consensus in the findings that cyberbullying is a stronger predictor of depression in adolescent girls than boys (Kowalski et al. 2014) and that girls report a stronger negative emotional impact of cyberbullying than boys (Ortega et al. 2009), highlighting the value of cyberbullying identification models that draw on critical insights from psychology research.

Some of the most informative findings from within psychology come from meta-analytic reviews, which reveal the average effect of various risk factors across multiple studies and different research teams, weighted by characteristics that contribute to the accuracy and overall quality of a specific research study (i.e., the size of the research sample). For instance, one of the most robust predictors of cyberbullying victimization among teens is whether they have also been victims of traditional bullying in the past (Kowalski et al. 2014; Guo 2016). In two separate meta-analyses (Kowalski et al. 2014; Guo 2016), previous history as a victim of traditional (i.e., offline, face-to-face) bullying emerged as the strongest of numerous risk factors for cyberbullying victimization. Furthermore, whereas symptoms of psychological distress and behavior problems are frequently examined as outcomes associated with cyberbullying victimization, the psychology literature also indicates that they are a robust *predictor* of cyberbullying victimization. Conclusions about the direction of causality between psychological distress, behavior problems, and cyberbullying cannot be made, given the correlational and largely cross-sectional nature of the data; in fact, it seems likely that a reciprocal relation exists, such that symptoms and indicators of poorer mental health increase adolescents' risk of being targeted by cyberbullies, which then contributes to increased psychological distress and the onset of new symptoms or negative behaviors. A crucial insight, however, is that adolescents' previous history of mental health and behavioral challenges are important factors to include in an identification model.

To begin incorporating these results into our automated identification model, we divided the set of bullying factors into *warning signs* (quantifiable measures like the number of insulting wall posts) and *vulnerability factors* (risk factors and circumstances that may increase the probability of experiencing cyberbullying). The current BullyBlocker model identifies warning signs and vulnerability factors by (1) analyzing the interaction of an adolescent with his or her network through wall posts and picture or video comments, (2) obtaining information from the adolescent's Facebook profile, like age, gender, and schools attended, and (3) obtaining relevant information about additional vulnerability factors directly from parent users of the BullyBlocker app. We mention next specific research findings that provide the framework for the current version of our model.

The current BullyBlocker identification model considers as warning signs: the number of insulting wall posts, the

number of embarrassing or insulting comments on photographs, and the number of embarrassing or insulting comments on videos that an adolescent has received in the last 90 days. (All posts or comments written by the potential victim are excluded.) In the absence of prior work examining the relative strength of various warning signs as indicators of cyberbullying, we use, whenever available, the correlation coefficients identified in meta-analytic reviews of cyberbullying research to compute the weights of the factors. When this is not possible, we include estimated weights that will be modified in subsequent models.

Data pertaining to vulnerability factors are collected in two ways. First, some information is extracted from the adolescent's social media profile. This includes, for example, the adolescent's age and gender. Nuances in the findings across studies concerning rates of cyberbullying victimization among different age groups and between males and females underscore the importance of including these demographic factors in our identification model, albeit with relatively lower assigned weights, to better understand how they may contribute to cyberbullying risk. Because cyberbullying victims are typically adolescents on the "fringe" of various peer groups (Piazza and Bering 2009), two factors that can contribute to a teen's fringe status—whether the teen has recently relocated to a new neighborhood or a new school—will also be mined from adolescents' Facebook profiles and included as vulnerability factors. Finally, when relevant information is provided, sexual orientation will be included as a vulnerability factor based on multiple studies indicating that non-heterosexual adolescents are more likely to experience cyberbullying than their heterosexual peers (Fedewa and Ahn 2011). In sum, social media data can be mined to identify the extent to which an adolescent possesses any of the vulnerability factors described above.

Additional information that parents can provide about their teens—by filling out a brief in-app "user profile" survey for the adolescent(s) they wish to monitor—will further increase the accuracy of the identification model by allowing us to include several additional vulnerability factors. Specifically, based on the two key meta-analyses described above (Kowalski et al. 2014; Guo 2016), parents will be asked about the frequency of their teen's internet use and whether, to their knowledge, there is any prior history of being bullied. Parents can also indicate in the user profile whether their teen has a known history of: (1) internalizing problems—symptoms of psychological distress that an individual directs inward, including anxiety, depression, and low self-esteem, and/or (2) externalizing problems—problem behaviors that are directed outward, toward an individual's environment, including disciplinary problems (e.g., suspension or expulsion from a school) and substance use. Averaging across multiple studies, these correlated yet distinct psychological factors have been identified as reliable predictors of cyberbullying victimization among adolescents (Kowalski et al. 2014; Guo 2016).

To summarize, the current BullyBlocker identification model considers as warning signs: the number of insulting wall posts, the number of embarrassing or insulting comments on photographs, and the number of embarrassing or insulting comments on videos. As vulnerability factors, the model considers: race, age, gender, sexual orientation, having recently moved to a new neighborhood or a new school, past bullying history, frequency of internet use, internalizing problems (depression, anxiety, and low self-esteem), and externalizing problems (disciplinary problems and substance abuse). The specific way in which the identified factors are used in our model is described in Sect. 3.

3 Automated identification of cyberbullying

The main design components of the BullyBlocker app are presented in Fig. 1. The app is designed for use by the parent or guardian of an adolescent, who will be required to enter the Facebook login information of the adolescent being monitored.

The *Data Collection Module* uses the adolescent's login information to retrieve all of the required information from Facebook, i.e., the stream of recent wall posts by other Facebook users, the streams of photograph and video comments made by other users, and user profile information such as age, gender, recently attended schools, and home location.

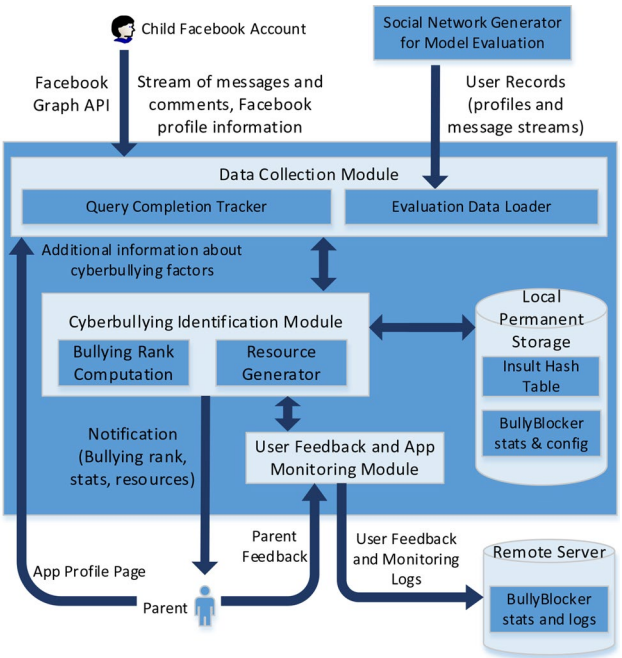

Fig. 1 BullyBlocker architecture

An important component of this module is the *Query Completion Tracker*, which keeps track of the asynchronous and parallel information requests sent to Facebook. The stream of messages retrieved from Facebook is organized using a tree structure of pages that contains different subsets of the root level messages, their comments, and sub-comments. The tracker guarantees that all of the pages associated with a given stream are properly traversed and processed. This module also makes sure that the frequency of requests sent to Facebook is under the threshold established by Facebook. In addition to obtaining the message/comment streams and the Facebook profile information, this module also collects information directly from the parent through the app using a brief survey, as shown in Fig. 3. This survey collects information related to additional vulnerability factors such as ethnicity, race, frequency of internet use, and previous bullying history.

To allow for the evaluation of the BullyBlocker identification model, we created the *Evaluation Data Loader*, a component of the Data Collection Module—used only during the evaluation process—to enable the loading of social network data from generated datasets instead of from actual users (see Sect. 4). The evaluation dataset is composed of user records, with each record containing a user profile and associated message stream. This dataset is generated by the *Social Network Generator* component using information from real-world cyberbullying interactions. The details of the Social Network Generator and the evaluation process are presented in Sect. 4.1.

The *Cyberbullying Identification Module* then uses the retrieved data to estimate the likelihood that an adolescent is a victim of cyberbullying on Facebook. To this end, the application computes a *Bullying Rank* expression, based on the identified warning signs and vulnerability factors, that aims to represent the probability that an adolescent is experiencing cyberbullying. The Bullying Rank is used to normalize the intensity of cyberbullying and to simplify the results presented to the parent. Figure 2 shows the general approach for computing the Bullying Rank.

As shown in Fig. 2, the Bullying Rank (BR) is computed based on the values of warning signs (WS) and vulnerability factors (VF). Each part is given an appropriate weight such that the range of the BR is [0, 100]. The Bullying Rank can fall into any of three pre-defined levels, with the respective intervals: low risk [0, 33], moderate risk [34, 66], and severe risk [67, 100].

The Bullying Rank, together with several aggregated measures such as the number of insulting wall posts, the number of insults in photograph and video comments, the number of potential bullies, and the time range of the analysis, is generated by the Cyberbullying Identification Module and loaded in the results page of the BullyBlocker app. This module also generates a customized list of resources,

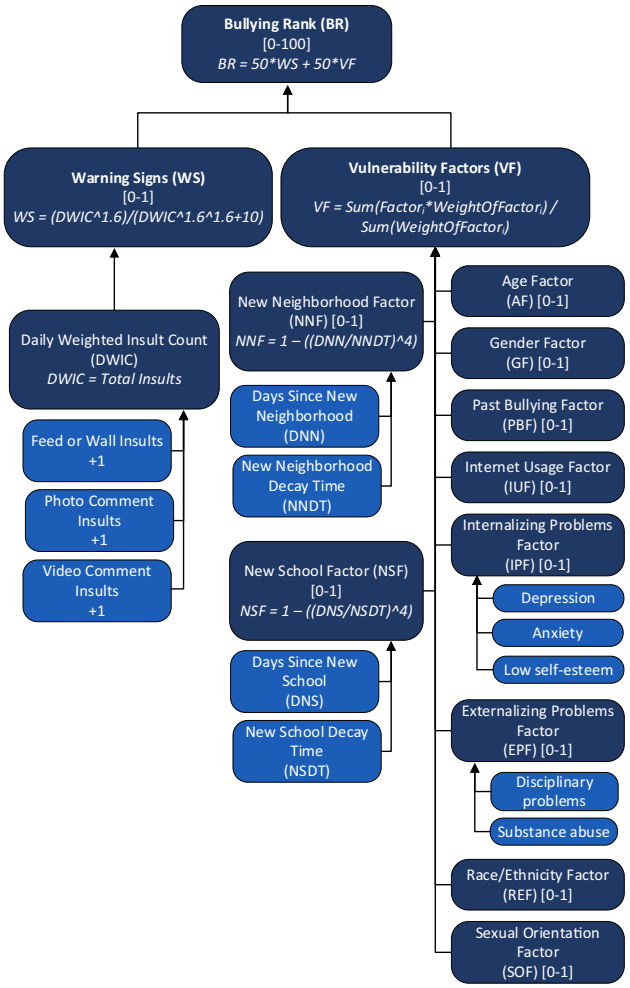

Fig. 2 Bullying Rank factors

including websites and hotlines, that direct parents to national and local organizations that provide information about ways to address current and prevent future instances of cyberbullying. Some of the information processed and generated by this module is stored in the mobile device's permanent storage. For instance, the app records the previously computed values of the Bullying Rank and its various components and the most recent dates on which the adolescent moved to a new neighborhood or school.

The *User Feedback and App Monitoring Module* enables parents to submit a brief survey about their perceptions of the accuracy of the app. This module also includes a monitoring component that can be enabled during the app test phase with a specific set of test users to log the computed Bullying Rank values, final values of the cyberbullying factors in Fig. 2, and an encrypted version of the user IDs. The collected data are sent to a remote web and database server for app monitoring and assessment purposes.

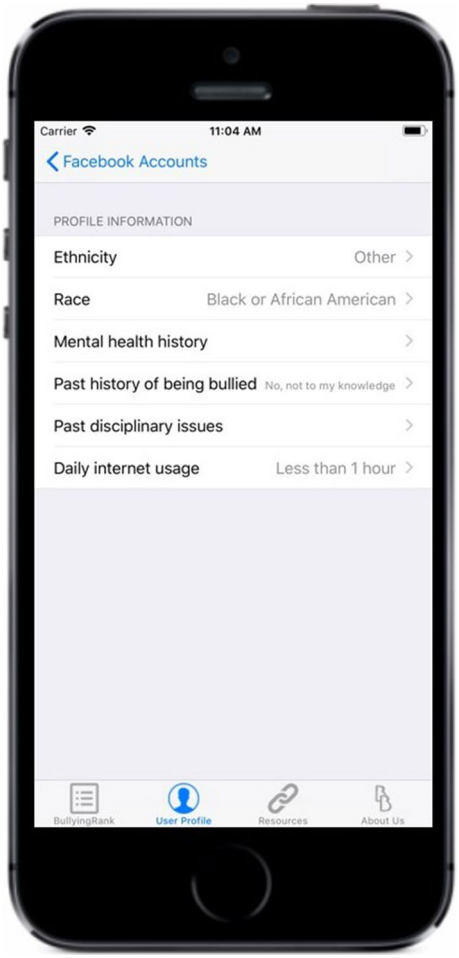

Fig. 3 User Profile Interface in the BullyBlocker app. *Note*: to avoid the use of psychological terms with which parents may be unfamiliar, the internalizing problems factor is referred to as "Mental health history" and the externalizing problems factor is referred to as "Past disciplinary issues" in the in-app user profile survey

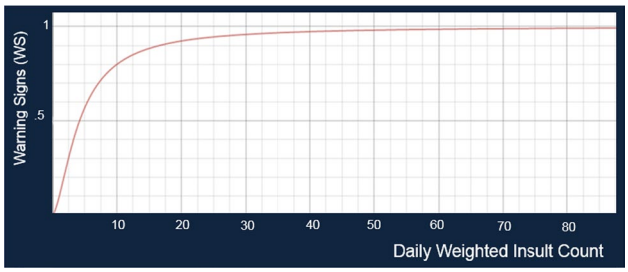

Fig. 4 Warning signs versus daily insult count

3.1 Measuring warning signs

The warning signs (WS) component aims to quantify the amount of insulting content received by the monitored adolescent. This component is included to account for the *Group Effect*, as identified in (Dooley et al. 2009), where the number of insults increases the severity of perceived victimization.

As shown in Fig. 2, this component is computed based on the number of feed (wall) insults, the number of insulting photograph comments, and the number of insulting video comments received by the adolescent during the last N days (currently $N=90$). To decide if a message is of insulting nature, we analyze the content in the message by running hash-based lookup operations on a dictionary of insults and their variations (variations are considered because, in many instances, adolescents use them instead of the original insulting words). The warning signs component could also consider the number of insulting private messages received by the potential victim when accessing these data is allowed by the social network's query API. This is currently not the case with Facebook. The raw insult counts are combined into the daily weighted insult count (DWIC) by applying equal weights to all sub-components (feed, photograph and video insults) and computing the average value per day. The DWIC value is then normalized to be in the [0, 1] range. Rather than applying uniform scaling, we use a function that assigns higher weights to initial insults, given that after a certain large value of daily insults (~30), additional insults tend to have a minimal effect. The function (specified in the warning signs box in Fig. 2) is plotted in Fig. 4. Observe that the X-axis of this graph corresponds to the values of DWIC and the Y-axis is the WS value computed using the equation previously referenced. As shown in this figure, going from 5 to 10 daily insults generates a larger increment in the function value than going from 80 to 85 insults. The current model uses 90 as the value of N (number of days). The meta-analysis of cyberbullying by Baldry et al. (Baldry et al. 2016) shows that this time frame is commonly used or contains the reference period used in previous psychology studies.

While all insults receive equal weight in the current model, the model can easily be extended to assign different weights to different types of insults and to increase the weight of an insult based on properties such as the number of people who "liked" the insulting message or the message's recency. We were not able, however, to identify studies that have directly addressed this aspect of cyberbullying.

3.2 Measuring vulnerability

The vulnerability factors (VF) component aims to quantify the level of vulnerability of the monitored adolescent. As shown in Fig. 2, this component is computed based on the following ten factors: age, gender, sexual orientation, days since transition to a new neighborhood, days since transition to a new school, race and ethnicity, prior history of being bullied, frequency of internet use, internalizing problems

(i.e., "mental health history" in the in-app survey), and externalizing problems (i.e., "past disciplinary issues" in the in-app survey). The value of each factor is in the range of [0, 1]. Figure 6 shows the details and weights of each factor. Intuitively, each factor should have a different weight in the identification model, given variability in the strength of the relation between each factor and cyberbullying risk. To capture this property, we have assigned weights primarily based on the correlation coefficients identified in previous comprehensive meta-analytic reviews. Specifically, the weights for age, prior history of being bullied, internalizing problems, and externalizing problems were based on the correlation coefficients identified in Kowalski et al. (2014) and Guo (2016). The weight for frequency of internet use was based on the correlation coefficient identified in Kowalski et al. (2014), and the weights for gender and race/ethnicity were based on the correlation coefficients identified in Guo (2016). The weight assigned to the sexual orientation factor was based on the meta-analytic effect in Fedewa and Ahn (2011), which was reported as an odds ratio indicating that gay, lesbian, and bisexual teens were 2.24 times as likely to be a victim of cyberbullying as heterosexual teens. We performed a transformation (see Bonnet 2007) to convert the odds ratio to a correlation coefficient. In the absence of published research syntheses examining the relation between cyberbullying risk and the new neighborhood and new school factors, these factors were assigned initial estimated weights in the current identification model.

For the age and gender factors (AF, GF), the model assigns a value of 1 when the age of the potential victim is between 11 and 16 years old, and the gender is female, respectively. The race/ethnicity factor (REF) is set to 1 if race is non-white or if ethnicity is Hispanic/Latino. The sexual orientation factor (SOF) is set to 1 when the potential victim is identified as LGBTQ (combining the gender and "interested in" properties of the Facebook profile). The past bullying factor (PBF) receives various values based on the recency of the previous history of being bullied. Similarly, the internet use factor (IUF) receives different values based on the frequency of internet use. The internalizing and externalizing problems factors (IPF, EPF) are assigned different values based on the number of sub-factors (listed in the corresponding rows of Fig. 6) identified in the in-app survey. The new neighborhood factor (NNF) and new school factor (NSF) weights are based on the number of days since the adolescent moved to a new neighborhood or school. The effect of these components is assumed to change over time, such that the effect should be higher if the adolescent moved recently. To represent this, the model uses a function, specified in the NNF and NSF boxes in Fig. 2 and plotted in Fig. 5, that generates a NNF or NSF value that starts at 1 (the day the adolescent moves to a new neighborhood or school) and decreases

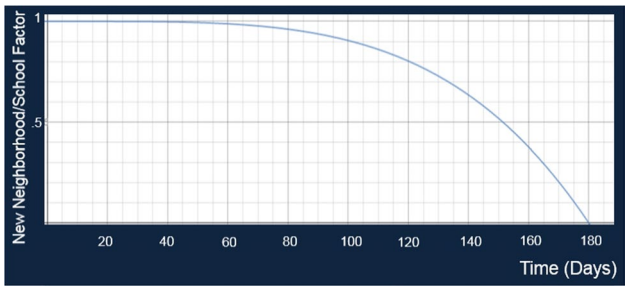

Fig. 5 New neighborhood/school factor versus time

Factor	Details	Weight
New School	# days in a new school	0.10
New Neighborhood	# days in a new neighborhood	0.10
Age	Applied if value is 11-16	0.04
Gender	Applied if value is female	0.12
Race/ Ethnicity	Applied if race is non-white or if ethnicity is Hispanic/Latino	0.02
Sexual Orientation	Applied if self-identified as LGBTQ	0.29
Past Bullying	Applied if user experienced bullying in last 1 month, 1-2 months, more than 2 months	0.42
Daily Internet Use	Considers ranges <1h, 1h-3h, 4h-6h, >6h	0.17
Internalizing Problems	Considers history of depression, anxiety, low self-esteem	0.28
Externalizing Problems	Considers history of disciplinary issues or substance use	0.21

Fig. 6 Details and weights of vulnerability factors

over time until it reaches a value of 0 (when the number of days is equal to a parameter value, e.g., new neighborhood decay time).

The total value of vulnerability factors is computed by multiplying each factor by its weight and then re-scaling the result to be in the range of [0, 1]. The design to compute the VF component can be extended in the future to integrate new empirical findings from both the psychology and computer science literatures. For instance, measures like the number of new friends added to an adolescent's social network since moving to a new neighborhood or school could be incorporated into the model.

Figures 7 and 8 show the interface of the current version of the BullyBlocker app. These figures show two screenshots that correspond to the result pages generated for two monitored adolescents, Maria and Constance Winther. While the Bullying Rank for Maria is relatively low (i.e., 6), the one for Constance is significantly higher (i.e., 83) due to much higher WS and VF values.

3.3 Smart generation of parent/victim resources

One of the challenges that parents face upon learning that their adolescent is a victim of cyberbullying is knowing how to respond effectively; that is, how to help curb the bullying attacks, prevent future instances of cyberbullying, and provide the necessary psychological and emotional support. Critical to an effective resolution is parents' ability to locate appropriate resources (e.g., anti-bullying organizations, hotlines, literature), yet finding the most relevant resources for a specific instance of cyberbullying can be an overwhelming task.

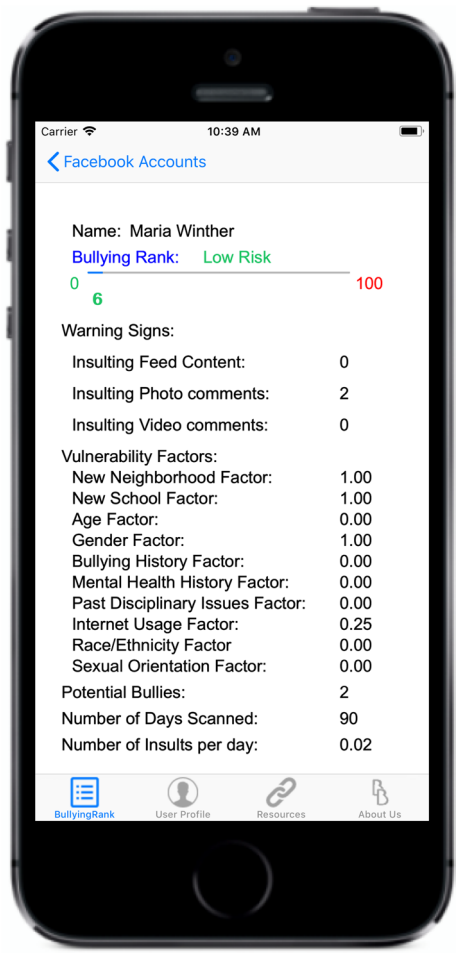

Fig. 7 BullyBlocker results—low risk

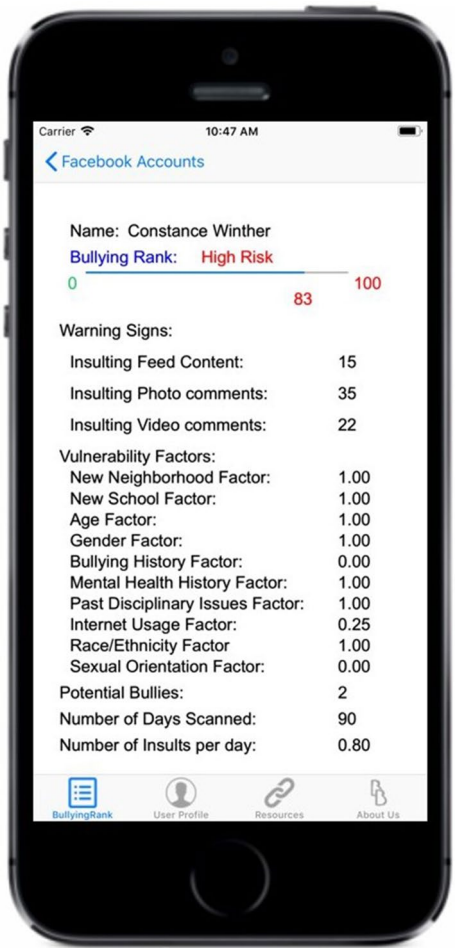

Fig. 8 BullyBlocker results—high risk

BullyBlocker aims to be an effective tool to address this problem by generating a customized list of resources that is tailored to the unique circumstances surrounding the bullying attack(s) and the individual needs of the adolescent. To this end, the app maintains an internal compact representation of the various factors that have been identified for the specific user being analyzed. The app also maintains a robust list of anti-bullying resources (local and national websites and hotlines) annotated with the specific groups targeted by each resource, e.g., racial and ethnic minorities, girls, members of the LGBTQ community. After completing the computation of the Bullying Rank, the app uses the information pertaining to the identified factors to rank the list of resources by potential relevance. As shown in Fig. 9, the list of most pertinent resources is presented at the top of the anti-bullying resources page. In this example, the vulnerability factors that were activated are race/ethnicity (the monitored adolescent identifies as Hispanic) and externalizing problems (history of substance use was reported). Considering this information, the app recommends a customized list of resources that includes links to the websites for Drug Rehab

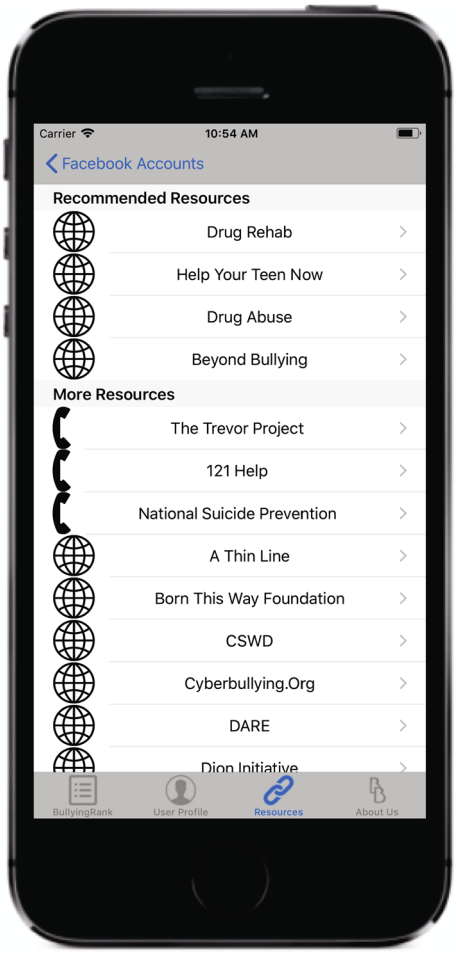

Fig. 9 BullyBlocker personalized anti-bullying resources

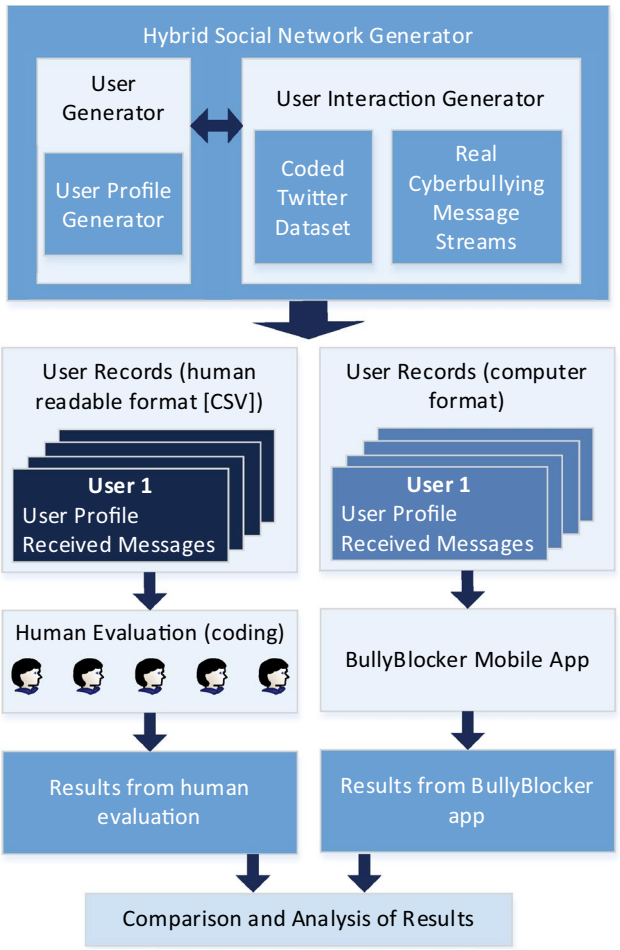

Fig. 10 Evaluation framework using a hybrid social network

(http://www.drugrehab.com/guides/bullying), Help your Teen Now (http://www.helpyourteennow.com/cyber-bullying-and-addiction-in-teenagers), and Drug Abuse (http://www.teens.drugabuse.gov/blog/post/four-things-know-about-cyberbullying); and race/ethnicity-based resources such as Beyond Bullying (http://wwwbeyondbullying.com/racistbullying.html).

4 Evaluation of holistic BullyBlocker identification models

As mentioned previously, one of the limitations of most of the prior work in this area is that the proposed models focus on the identification of cyberbullying in a single post, message, or picture. In many real-world cases, however, cyberbullying involves a repeated sequence of insults, or insult bursts (Hinduja and Patchin 2013; Squicciarini et al. 2015). To our knowledge, our model is one of the first that seeks to identify cyberbullying by considering multiple streams of messages (e.g., wall posts, picture/video comments). Furthermore, our model integrates a set of vulnerability factors based on extensive empirical work in psychology and related social science fields. Finally, the evaluation of holistic models like the one presented in this paper requires a more comprehensive evaluation framework than those needed for simpler models. Yet, a key challenge stems from the difficulty of obtaining datasets that contain all of the required information. That is, whereas the generation and labeling of datasets is relatively simple with models that aim to predict cyberbullying in a single message, the complexity increases when considering message streams and multiple vulnerability factors.

To address these challenges, we propose the evaluation framework for holistic cyberbullying identification models depicted in Fig. 10. The goal of this framework is to identify whether or not cyberbullying has occurred by considering the entire user profile as well as the streams of messages received by the user. In this section, we discuss the proposed evaluation framework and its underlying hybrid social network in detail, and present the results obtained

using this framework to evaluate the BullyBlocker identification model.

4.1 Evaluation framework

The proposed evaluation framework uses a hybrid social network generator to create realistic datasets that are provided as input into the app and also used later in the human evaluation phase.

Hybrid social network generator This component of the framework generates a test social network composed of synthetic users and real-world interactions (i.e., messages).

The *User Generator* module outputs a set of N users and their profile information. The profile information consists of the attributes associated with the vulnerability factors presented in Fig. 6 (e.g., age, gender, race, frequency of daily internet use, bullying history). With the exception of the bullying history attribute, specific values for the different vulnerability factors were evenly distributed among all possible values or ranges. For the previous bullying history attribute, the distribution of values was: no previous bullying (50%), experienced bullying last month (16.66%), from one to 2 months (16.66%), and more than 2 months ago (16.66%).

The *User Interaction Generator* module produces a set of interactions (message streams) among the created users. The goal of this module is to create message sequences that are similar to the ones found in real-world social networks. To this end, this component uses two sources of real-world messages: (1) a coded (labeled) Twitter dataset (composed of subsets of cyberbullying and non-cyberbullying (i.e., normal) messages), and (2) real cyberbullying message streams. The coded Twitter dataset was obtained following the procedure suggested in (Nand et al. 2016).

We crawled this dataset using the Twitter streaming API (Morstatter et al. 2013) from September 19th to 25th, 2017 with the following keywords: *nerd, gay, loser, freak, emo, whale, pig, fat, wannabe, poser, whore, should, die, slept, caught, suck, slut, live, afraid, fight, pussy, cunt, kill, dick, bitch.* We initially obtained 4,730,766 tweets. After the initial data collection, we employed the pre-trained *Bully* classifier (Xu et al. 2012) to label each tweet in the crawled dataset and extracted a refined subset with high confidence, containing 7500 positive samples (i.e., cyberbullying) and 7500 negative samples (i.e., normal). Then, these 15,000 tweets were further labeled by two well-trained human annotators with backgrounds in psychology and computer science. A third trained annotator was asked to resolve any discrepancies between the ratings of the initial two annotators. After resolution of discrepancies and data cleaning, we obtained the final dataset composed of 3647 cyberbullying tweets (referred to as TwitterCB) and 11,347 normal ones (referred to as TwitterNonCB). The second dataset (NewsCB) was composed of real cyberbullying message streams found in real social networks, e.g., Facebook, Twitter and Instagram. Many of these streams were obtained from news articles reporting well-known instances of cyberbullying. Due to required manual work to identify and extract these cyberbullying sequences, their number is relatively small (100 streams, where each stream contains between one and eleven messages). An important benefit of this dataset, however, is that it captures information about the way cyberbullying messages are distributed over time. These temporal properties are maintained in the message streams generated using this data source.

The User Interaction Generator module uses both message sources to generate the interactions among the created users. The key parameters used in this step are the total number of users ($N = 400$), the number of cyberbullying streams in NewsCB ($K = 100$), the number of days ($D = 90$), and the maximum number of messages per user ($M = 100$). The messages were generated as follows: For each of the first K users, the stream of the ith user contains all of the cyberbullying messages included in the ith stream in NewsCB. The remaining messages (totaling M messages per stream) are generated adding BF*(M − NewsCB[i].length) cyberbullying messages and (1 − BF)*(M − NewsCB[i].length) normal messages. BF (bullying fraction) is the fraction of the remaining messages that are cyberbullying interactions. This value, in the range [0.0, 1.0], is randomly computed for each of the first K users. For each of the remaining N–K users, the message stream of a given user is generated by interleaving BFF*D bullying messages from TwitterCB and (M-BFF*D) normal messages from TwitterNonCB. Bullying frequency factor (BFF) is the number of cyberbullying messages that a user receives per day and is also randomly generated for each of the remaining users in the range [0.0, 1.0].

We expect that the hybrid network generator and the real-world datasets will be used by other researchers to evaluate the performance of future comprehensive and holistic identification models. To facilitate these tasks, we have made available the source code of the generator and its input datasets (https://bullyblocker.project.asu.edu/data).

Generated datasets A final step of the hybrid social network generator is to produce two datasets capturing the information of the social network. In both datasets, each record represents the social network data associated with a single user. Each user record is composed of (1) the user profile information (user ID and vulnerability related features) and (2) the set of messages received by this user. Each message contains the ID of the user sending the message, the timestamp, and the message content.

The two datasets contain the same data but use different representation formats. Each record in the first dataset uses a human-readable format. Each record in the second dataset is structured as a document intended to be processed programmatically using the BullyBlocker app. The only content difference between the two datasets is that the human-readable one includes a flag that identifies the cyberbullying-related messages. A sample of a produced user record is presented in Fig. 11.

Human evaluation The generated human-readable dataset containing the information of the hybrid social network was evaluated by members of our research team. Each record was assessed independently by two designated annotators, and any discrepancies between the annotators' assessments were resolved by a third rater. Evaluating a user record entailed assessing the user's entire profile information and stream of messages, and assigning the record a Bullying Rank between 0 and 100 to reflect the probability that the user is experiencing or has recently experienced cyberbullying. A discrepancy in annotators' assessments was defined as two people assigning Bullying Rank values in different risk level categories [low risk (0, 33), moderate risk (34, 66), and severe risk (67, 100)].

Evaluation using the BullyBlocker app The generated user records were also processed by the BullyBlocker app. To this end, the app was extended by a module that read from the generated dataset instead of obtaining the information from Facebook. This module also executed the Bullying Rank Computation task for each user and saved the Bullying Rank values generated by the app.

Comparison and analysis of results In the last step of the process, we compared the results obtained from the human evaluation phase against the results obtained using the automated BullyBlocker cyberbullying identification model. The results of this comparison are presented in the next subsection.

Name: User0
Age: 17
Gender: Female
Ethnicity: Other
Race: African_American
Depression: No
Anxiety: No
Self Esteem Issues: No
History Bullied: The user was bullied last month
Disciplinary Issues: Yes
Substance Abuse: No
Internet Usage: Weekly use between 4 and 6 hours

Timestamp	SenderID	Bullying	Text
11/19/2017 5:19	102	B	No he sounds like a dick https://t.co/Dptnt0E3H0
11/20/2017 2:09	71		If you don't cry yourself to sleep every night to miserable at best, are you even emo?
11/19/2017 11:15	102	B	Calling is bad. If it wasn't for that pass interference wouldn't of scored. He suck bro
11/20/2017 22:06	139	B	Can't sing about someone's dick but you can call Marcos Alonso a murderer at the m
11/20/2017 5:16	102	B	You should eat!
11/22/2017 4:58	44		"If you see a good fight, get in it."
11/19/2017 17:31	16	B	TF???? EAT SOMETHING!!!!!!!!!!!!!
11/25/2017 3:26	107	B	Those Are Bones. Need to eat
11/22/2017 5:09	39	B	I hope your 99 kids are fine
12/12/2017 20:52	140	B	I cant beleave that fat poo ball owns the goodlooking girl contest and they say he w
11/21/2017 23:51	21	B	Word this nigga is dick eatin hard mf tags me in a post EVERYTIME the giants lose li
11/25/2017 9:05	319		See.... because I will fight you right now!! _Ã¯Ã–Ã
11/23/2017 22:46	91	B	Too skinny
11/21/2017 0:45	39	B	Her skin is unique and beautiful, but she seems too thin.
11/28/2017 16:27	153	B	Don't hesitate to give your opinions either, you gay. I'll read them in the mornin
12/4/2017 11:49	114		#RETWEET IF YOU LIVE IN #ENGLAND AND FOLLOW ME FOR #FOLLOWBACK !
12/8/2017 3:45	67	B	Whoever fucked me over on 2015, I hate you bitch https://t.co/ai2aOOCx4P
12/9/2017 22:50	367		Hour and a half, that's close-ish! I'd love to live that close to a sea or an ocean
12/12/2017 3:56	16		love to crack open a bundaberg with die jungen
12/14/2017 10:20	87	B	waw..not only slutty ..v...i feel that you are slut ä__ôÍ__ôÍ__ôÎì
1/2/2018 2:02	124	B	bitch get over that heartbreak and keep it tf movin
11/25/2017 4:07	246		Yes Oh.The boy dey kill us with off key here
11/22/2017 18:46	67	B	Too thin
12/13/2017 17:46	32		youâó»re now the reason that I fight
12/11/2017 20:56	389		Only 30 minutes until I'm live on twitch with more Earthbound! https://t.co/uz4jmMSI
12/9/2017 15:34	202		But his point is lost? Troops win battles, and if they don't fight, they don't win. They
12/2/2017 20:44	124	B	Love you Joel but I am weary fighting Thyroid #Cancer & feeling there is nobod
11/24/2017 1:42	43	B	Eat few burritos
1/12/2018 11:33	22	B	Get that son of a bitch out of the White House.
12/25/2017 15:34	151	B	because i wil fucking die if i see one i swear i'll kill everyone in the cinema
12/16/2017 20:54	107	B	That's absolute bullshit. Our troops fight for the right to protest. They defend the co
12/5/2017 21:41	96		KILL IT GIRL YES SLAY
12/23/2017 13:17	183		Why does Dysphoria make me afraid to talk about girls like I am one, why does my b

Fig. 11 Sample user record

4.2 Evaluation results

Figure 12 shows the average error of the Bullying Rank values computed by the proposed BullyBlocker model, with the Bullying Rank (probability that an adolescent is being cyberbullied) expressed as a percentage value (1–100). We compute the error as the absolute value of the difference between the Bullying Rank produced by the app and the average value of the human coding results. This figure presents the average error for various weight configurations of the two main components of the Bullying Rank (warning signs and vulnerability factors). As shown in Fig. 12, the BullyBlocker app produces the smallest error (18.4 percentage points) when there is an even distribution of weights between warning signs and vulnerability factors (50–50%). The error gradually increases as either of the components is weighted more heavily than the other. Based on these results, we set the weights of both components to 50% in the latest version of the app.

Figure 13 presents the frequency of errors for various error ranges. The results show that when the weights of warning signs and vulnerability factors in the BullyBlocker

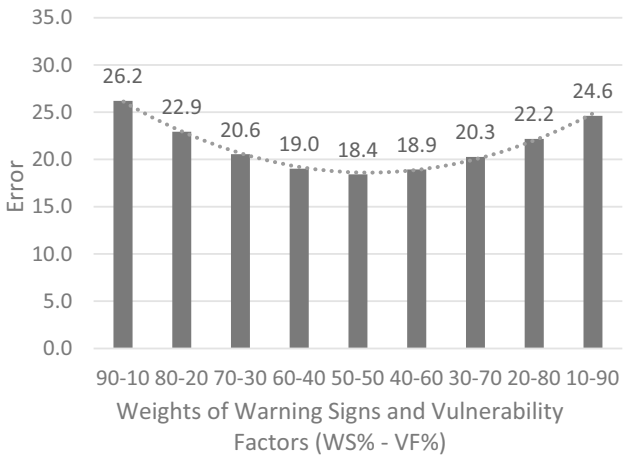

Fig. 12 Average error for different weights of warning signs and vulnerability factors

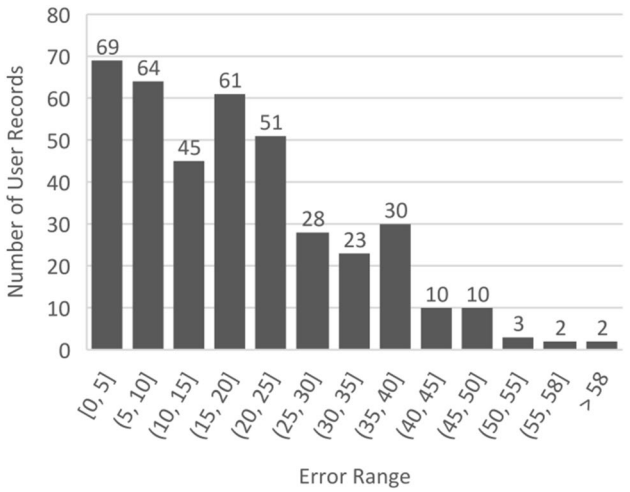

Fig. 13 Histogram of errors

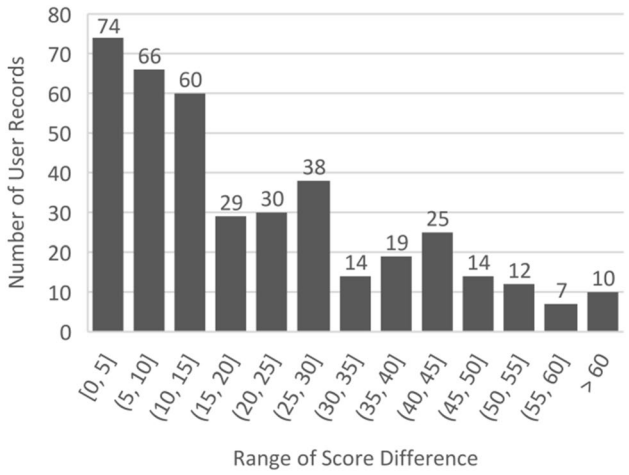

Fig. 14 Histogram of the score difference between human coders

app were set to 50%, the most frequent error values fell within the lowest ranges. That is, in 33% of the cases, the error (reflecting the difference between the Bullying Rank estimated by the human annotators and the Bullying Rank calculated by the app) was smaller than 10 percentage points, while in 60% of the cases, the error was smaller than 20 percentage points.

The results in Figs. 12 and 13 should be interpreted considering an inherent challenge that human annotators face when estimating the probability that an adolescent is being cyberbullied, which stems from subjectivity in how annotators approach the task of quantifying cyberbullying risk. That is, we had eight human annotators each rate a subset of 100 cases from the human-readable version of the dataset; with each of the 400 cases generated by the hybrid social network rated by two annotators independently. Whereas annotators maintained a consistent strategy for estimating cyberbullying risk for the 100 cases to which they were assigned, individual differences in how each annotator interpreted the data holistically and translated their assessments into the numerical Bullying Rank index likely emerged. Figure 14 shows the distribution of the score difference in the Bullying Rank estimates made by the two annotators who evaluated each case during the human evaluation phase. The distribution in this figure is similar to the distribution of the BullyBlocker model error presented in Fig. 13. Specifically, in 35% of the cases, the score difference between the two human annotators assigned to a particular case was smaller than 10 percentage points, while in 57% of the cases, the score difference was smaller than 20 percentage points.

The results presented in this section show that the proposed model produces relatively small error in most cases. We expect that the error levels of the model will be further reduced by integrating some of the techniques described in the future work section.

5 Future work

As an emerging sphere of research, efforts to develop and evaluate the accuracy of models for the automated identification of cyberbullying can benefit from future work in several key areas. In this section, we describe some of these areas and provide details of our team's progress along these research paths.

5.1 Alternative identification models

This area involves the use of other computational techniques like similarity-aware data processing and machine learning to build alternative holistic cyberbullying identification models that consider both an array of profile information features and the users' streams of messages. Two important tasks in this area are the comparison of multiple models and the study of integration mechanisms to build highly accurate hybrid models.

Similarity-aware model for cyberbullying identification To this end, we are investigating cyberbullying identification models that use the power of similarity operators (Tang et al. 2016; Silva et al. 2013; Yu et al. 2007; Silva et al. 2015), i.e., data processing operators like the Similarity Join and Similarity Grouping that identify and exploit similarities in the data. An initial idea to build this model is to use a vector-based representation of a person's behavior. Our efforts in this area are directed toward building wide feature vectors using the information on cyberbullying risk factors identified in our current BullyBlocker model. Specifically, the vector could include numeric measures for various warning signs (e.g., number of insulting messages) and vulnerability factors (e.g., a recent move to a new neighborhood or school). For example, considering an initial set of factors,

the structure of the vector would be as follows: #EmbarrassingPictures, #TotalFeed Messages, #TotalPictureComments, #Pictures, #Bullies, #Friends, …, Age, Female?, #DaysSinceNewNeighborhood, #DaysSinceNewSchool, Hispanic?, AfricanAmerican?, …). Moreover, we are exploring the inclusion of factors that are associated with the role and position of the potential victim in his or her social network [by analyzing features like closeness centrality, betweenness centrality, degree centrality, eigenvector centrality, and clustering coefficient (Golbeck 2013)], as well as features aimed at capturing language patterns based on the message streams. This model will also represent common cyberbullying behavior patterns as cyberbullying behavior vectors. The outcome of this sub-task will be a number of cyberbullying feature vectors that, as a group, represent the most common patterns of cyberbullying victimization. The distance of an adolescent's feature vector to the cyberbullying behavior vectors can be used to estimate the likelihood that a person is being cyberbullied. This approach also enables other interesting types of analyses. For instance, using clustering or Similarity Grouping, we can identify and study groups of adolescents who are experiencing similar types of social, emotional, or behavioral issues.

Machine learning model for cyberbullying identification We are also working to design and study comprehensive machine learning models (Kelleher et al. 2015) for cyberbullying identification. The cyberbullying identification problem can be modeled as a classification (discrete output classes, e.g., low, moderate, and high cyberbullying risks) or regression (continuous output, e.g., Bullying Rank value) task, and multiple strategies can be used to implement them, e.g., Logistic Regression, Sparse Neural Networks, Support Vector Machines, and Naïve Bayes. Specifically, we are in the early stages of designing a Sparse Neural Networks model. This approach will enable building a global artificial neural network by connecting smaller complete neural networks that can focus on specific classification sub-tasks, e.g., considering subsets of the warning signs and vulnerability factors. This approach will allow us to build the global model incrementally by designing, implementing, and training individual neural networks that consider well-defined subsets of the cyberbullying factors (e.g., factors pertaining to race and gender). Some of these small neural networks may be based, in fact, on previously proposed models (Dinakar et al. 2011; Rafiq et al. 2015; Hosseinmardi et al. 2016; Reynolds et al. 2011; Huang et al. 2014; Squicciarini et al. 2015).

5.2 Integrating new vulnerability factors

Another important area of future work is the integration of new vulnerability factors into cyberbullying identification models like the one presented in this paper. In this area, we plan to continue drawing on emerging research findings in psychology to guide the identification of additional factors. Among the factors we plan to integrate are physical stature/weight, disability status, and concurrent use of multiple social networking sites. Other factors that, to our knowledge, have yet to receive empirical attention pertain to an adolescent's minority status *within* their specific school environment, neighborhood, or community. Similarly, socioeconomic status, religious identity, and immigrant status—and, importantly, the extent to which these aspects of a teen's identity contribute to their minority or fringe status within their immediate social environment—may also provide valuable insights for the identification of cyberbullying risk. Moreover, variables like degree of parental oversight of social media use and limitations on an adolescent's access to technology or social media can be modeled as protective factors associated with a decreased likelihood of being cyberbullied. These factors have been explored in a small handful of previous studies (Kowalski et al. 2014), although additional research is needed to better understand the extent to which they might buffer cyberbullying risk. Other information collected through the app, such as changes in one's relationship status, deletions from one's friend list, and hiding certain posts from view on one's newsfeed or timeline in Facebook, may provide an unprecedented mechanism for tracking meaningful changes in one's peer circle. Interestingly, most of the psychology research on cyberbullying has relied on adolescents' self-reports (Kowalski et al. 2014; Tokunaga 2010) of victimization, which may be influenced by their reluctance to report instances of cyberbullying as well as other methodological limitations (Kowalski et al. 2014). Data collected through the BullyBlocker app can thus circumvent several issues stemming from self-report measures of cyberbullying.

5.3 Expanding the synergy with the psychology community

Applications like BullyBlocker can also help inform the work of psychologists from both a research and clinical practice perspective. For example, parent feedback regarding the perceived accuracy of the app's underlying identification model can be compared and combined with clinical experts' assessments of cyberbullying, yielding a promising avenue for future psychological research. Future studies could, for instance, examine the degree of overlap in clinicians' and parents' assessments of cyberbullying, and compare each with adolescents' self-reports. Feedback from clinicians and parents will also be beneficial for understanding the extent to which various ranges of Bullying Rank values map onto the presence and severity of clinical symptoms that are directly observed by parents and clinical experts. Furthermore, parent feedback could also provide a platform for

investigating parents' more general attitudes about the use of automated tools for identifying a broad range of behavioral issues. Automated tools also have the potential to aid in the identification of symptoms of depression and anxiety, undue amounts of stress, low self-esteem, relationship violence, indicators of self-harm, and suicidal thoughts. Feedback from parents and adolescents can provide essential usability information (e.g., what level of detail about identified cyberbullying instances parents feel most comfortable receiving through the app, what level of detail might deter adolescents from providing their parents with their social networking site login information).

6 Conclusions

Cyberbullying is the most common online risk for adolescents. While the prevalence and determinants of cyberbullying have received considerable attention among researchers in the psychology community, there has been relatively little work on the automatic identification of cyberbullying in social networking sites, and even less work that seeks to bridge the efforts from computer science and psychology. This paper thus proposes a computational model for cyberbullying identification that builds on the research findings within the psychology literature. The paper also describes the design of BullyBlocker, an app that implements the proposed model, discusses the model's effectiveness in the context of a newly developed evaluative framework, and presents several ways in which the model can be extended. Our hope is that BullyBlocker, which has been recently made available through the Apple App Store, will have a strong societal impact, by identifying youth most vulnerable to cyberbullying victimization and by enabling parents to help their children in time to make a difference.

Acknowledgement The authors would like to thank ASU students Lisa Tsosie, Jaime Chon, Tara Tucker, Chance Brown, Liz Garcia, Bryan Sawkins, Rusty Conway, Anthony Nieuwenhuyse, Tom Schenk, Lu Cheng, Ashley Trow, Ayush Sanyal, Linle Jiang, Victoria Delgadillo, and Carmen Sanchez for their contributions to the design and implementation of the BullyBlocker app.

References

Baldry AC, Farrington DP, Sorrentino A (2016) Cyberbullying in youth: a pattern of disruptive behaviour. Psicología Educativa 22(1):19–26

Beyond Bullying Website. http://wwwbeyondbullying.com/racistbullying.html

Bonnet DG (2007) Transforming odds ratios into correlations for meta-analytic research. Am Psychol 62(3):254–255

BullyBlocker App in the Apple Store. Apphttps://itunes.apple.com/us/app/bullyblocker-app/id1236410370?mt=8

BullyBlocker Project Website. https://bullyblocker.project.asu.edu/data

Dinakar K, Reichart R, Lieberman H (2011) Modeling the Detection of Textual Cyberbullying. In: The international AAAI conference on web and social media (ICWSM)

Dooley JJ, Pyżalski J, Cross D (2009) Cyberbullying versus face-to-face bullying: a theoretical and conceptual review. J Psychol 217(4):182–188

Drug Abuse Website. http://www.teens.drugabuse.gov/blog/post/four-things-know-about-cyberbullying

Drug Rehab Website. http://www.drugrehab.com/guides/bullying

Fedewa AL, Ahn S (2011) The effects of bullying and peer victimization on sexual-minority and heterosexual youths: a quantitative meta-analysis of the literature. J GLBT Family Stud 7:398–418

Golbeck J (2013) Analyzing the social web. Morgan Kaufmann, Burlington

Guo S (2016) A meta-analysis of the predictors of cyberbullying perpetration and victimization. Psychol Sch 53(4):432–453

Help your Teen Now Website. http://www.helpyourteennow.com/cyber-bullying-and-addiction-in-teenagers

Hinduja S, Patchin JW (2008) Cyberbullying: an exploratory analysis of factors related to offending and victimization. Deviant Behavior 29(2):129–156

Hinduja S, Patchin JW (2013) Social influences on cyberbullying behaviors among middle and high school students. J Youth Adolesc 42(5):711–722

Hosseinmardi H, Rafiq RI, Han R, Lv Q, Mishra S (2016) Prediction of cyberbullying incidents in a media-based social network. In: The IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)

http://www.bullyingstatistics.org/content/cyber-bullying-statistics.html

http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/

Huang Q, Singh VK, Atrey PK (2014) Cyber bullying detection using social and textual analysis. In: The 3rd international workshop on socially-aware multimedia (SAM)

Kelleher JD, Namee BM, D'Arcy A (2015) Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies, 1st edn. The MIT Press, Cambridge

Kowalski RM, Limber SP (2007) Electronic bullying among middle school students. J Adolesc Health 41(6):S22–S30

Kowalski RK, Giumetti GW, Schroeder AN, Lattanner MR (2014) Bullying in the digital age: a critical review and meta-analysis of cyberbullying research among youth. Psychol Bull 140(4):1073–1137

Mishna F, Khoury-Kassabri M, Gadalla T, Daciuk J (2012) Risk factors for involvement in cyber bullying: victims, bullies and bully–victims. Child Youth Serv Rev 34(1):63–70

Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? comparing data from Twitter's streaming API with Twitter's Firehose. In: Proceedings of the 7th international AAAI conference on web and social media (ICWSM)

Nand P, Perera R, Kasture A (2016) How bullying is this message?: a psychometric thermometer for bullying. In: COLING, pp 695–706

Ortega R, Elipe P, Mora-Merchin JA, Calmaestra J, Vega E (2009) The emotional impact on victims of traditional bullying and cyberbullying: a study of Spanish adolescents. J Psychol 217(4):197–204

Patchin JW, Hinduja S (2012) Cyberbullying—an update and synthesis of the research. In: Patchin JW, Hindija S (eds) Cyberbullying prevention and response. Routledge, London, pp 13–35

Piazza J, Bering JM (2009) Evolutionary cyber-psychology: applying an evolutionary framework to Internet behavior. Comput Hum Behav 25(6):1258–1269

Rafiq RI, Hosseinmardi H, Han R, Lv Q, Mishra S, Arredondo-Mattson S (2015) Careful what you share in six seconds: detecting cyberbullying instances in Vine. In: The IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)

Reynolds K, Kontostathis A, Edwards L (2011) Using Machine Learning to Detect Cyberbullying. In: The 10th international conference on machine learning and applications and workshops (ICMLA)

Silva YN, Pearson S, Cheney JA (2013) Database Similarity Join for Metric Spaces. The International Conference on Similarity Search and Applications (SISAP), vol 8199. Springer LNCS, pp 266–279

Silva YN, Pearson SS, Chon J, Roberts R (2015) Similarity Joins: Their implementation and interactions with other database operators. Inf Syst 52:149–162

Silva YN, Rich C, Hall D (2016) BullyBlocker: towards the identification of cyberbullying in social networking sites. In: The IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), 2016, pp 1377–1379

Squicciarini A, Rajtmajer S, Liu Y, Griffin C (2015) Identification and characterization of cyberbullying dynamics in an online social network. In: The IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)

Tokunaga RS (2010) Following you home from school: a critical review and synthesis of research on cyberbullying victimization. Comput Hum Behav 26(3):277–287

Van Geel M, Vedder P, Tanilon J (2014) Relationship between peer victimization, cyberbullying, and suicide in children and adolescents: a meta-analysis. JAMA Pediatr 168(5):435–442

Williams KR, Guerra NG (2007) Prevalence and predictors of internet bullying. J Adolesc Health 41(6):S14–S21

Wolke D, Lereya T, Tippett N (2016) Individual and social determinants of bullying and cyberbullying. In: Vollink T, Dehue F, Guckin C (eds) Cyberbullying. Routledge, London, pp 26–53

Xu J-M, Jun K-S, Zhu X, Bellmore A (2012) Learning from bullying traces in social media. In: Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, pp 656–666

Ybarra ML, Mitchell KJ (2008) How risky are social networking sites? a comparison of places online where youth sexual solicitation and harassment occurs. Pediatrics 121(2):e350–e357

Tang M, Tahboub RY, Aref WG, Atallah MJ, Malluhi QM, Ouzzani M, Silva YN (2016) Similarity Group-by Operators for Multi-dimensional Relational Data. IEEE Trans Knowl Data Eng 28(2):510–523

Yu C, Cui B, Wang S, Su J (2007) Efficient index-based knn join processing for high-dimensional data. Inf Softw Technol 49:332–344